ChinaXiv合作期刊 图书情报工作
第62卷 第2期 2018年1月

科技论文的研究设计指纹自动识别方法构建与实现*

■ 钱力[1] 张晓林[1] 王茜[2]
[1] 中国科学院文献情报中心 北京 100190 [2] 中国医学科学院医学信息研究所图书馆 北京 100005

摘要：［目的/意义］从科技论文中自动识别与抽取研究设计指纹，能够为科研人员项目设计、研究方法的有效性评估、研究过程问题诊断、研究结果鉴别与评价提供重要的方法论和研究操作支撑。［方法/过程］基于科技论文研究设计指纹的概念模型，提出基于多规则模式混合机器学习方法，设计并实现指纹识别算法，并以数据挖掘领域的期刊文献数据为例，对识别算法的可行性与有效性进行分析验证。［结果/结论］除研究数据与研究趋势外，其他研究设计指纹识别准确率的认可度都基本达到80%以上，覆盖率的认可度，除研究工具与研究数据外，基本达到80%以上。

关键词：研究设计指纹 语义标注知识抽取 机器学习
分类号：TP391 G25
DOI：10.13266/j.issn.0252-3116.2018.02.018

1 引言

科技论文作为科学技术发展的重要战略资源，记录着科学真理验证过程、实验观测结果及研究结论等研究知识脉络线索，论文中所涉及的研究设计（包括研究问题、研究方法、研究流程、研究工具、相关方法与技术参数设定等），为后续研究者提供了宝贵的方法论和研究操作基础，成为科研人员项目设计、研究方法有效性评估、研究过程问题诊断、研究结果鉴别与评价的重要基础。科研人员希望能够有工具来有效回答“有谁用什么方法来解决这个问题”“哪些方法及其技术与参数设定能够更好地解决这个问题”等。但在科研文献数量迅速增加的环境下，在项目策划、设计、申请、立项、实施细节规划、实施管理等各个阶段，研究人员需要能够及时、准确地发现针对研究问题的各类研究设计及其细节，系统比较同一问题上不同研究设计及其成效，利用已有的各类研究设计及其执行效果来优化或调整自己的设计及研究过程，提供支持相应研究方法及其细节设置的知识证据链，而目前以主题词为主的数据挖掘或者聚焦于文摘层面的知识发现理论与技术还很难有效满足这些需求。

因此，设计并实现一套自动识别与抽取论文研究设计指纹的理论与技术方法体系就变得十分必要与迫切。笔者在已构建完成科技论文研究设计指纹概念模型[1]和识别模型的基础上，进一步研究与探索科技论文研究设计指纹自动识别的方法与实现。本研究结构如下：①界定研究设计指纹内涵与特征；②综述研究设计指纹自动识别方法相关的方法；③面对问题，提出并设计研究设计指纹自动识别方法；④结合实验数据验证研究设计指纹自动识别方法的可行性与有效性。

2 研究设计指纹内涵与特征

研究设计指纹是在一篇科技论文中能够唯一表示与描述科学研究设计的各个研究阶段与研究实体的重要知识单元，包括研究假说、研究目的、研究背景、研究方法、研究数据、研究工具、研究结果、研究结论以及研究趋势9种指纹类型，具备4个主要特征：①知识唯一性，即这些重要知识单元在遵守科研道德规范的前提下，其所具有的研究设计指纹特征是唯一的，其特征的核心构成维度有作者与文章标题；②研究思维性，即研究设计指纹可以精炼地揭示“一个科学研究设计的整体设计思路”；③知识结构性，即研究设计指纹可以结构化地描述“科学研究方法、过程和结果”，将其中的重要知识进行抽取、组织与关联；④骨干网络性，即一

* 本文系中国科学院文献情报能力建设专项“科技论文的研究设计指纹自动语义标注工具研发”（项目编号：院1658）研究成果之一。
作者简介：钱力（ORCID：0000-0002-0931-2882），信息系统与知识计算中心主任，副研究馆员，硕士生导师，Email：qianl@mail.las.ac.cn；张晓林（ORCID：0000-0001-8891-8366），研究员，博士生导师；王茜（ORCID：0000-0002-8629-8199），馆员，博士。
收稿日期：2017-08-30 修回日期：2017-11-14 本文起止页码：135-143 本文责任编辑：王传清

篇科技论文利用研究设计指纹可以类似于网络骨干图一样,可视化地描绘“科学研究中的骨干知识”。

3 相关研究方法

研究设计指纹作为蕴藏在科技论文内容中的特殊语义标签,与研究设计指纹识别相关的方法主要类似于语义标注与识别的知识抽取方法,以计算机程序自动执行的模式,实现研究方法等具有语义的知识构件自动识别与抽取。

3.1 基于本体知识工程进行识别知识的方法

R. Girju 等[2]等利用该方法自动识别英文句子中名词之间的语义关系(原因 - 结果、产品 - 生产者、内容 - 容器、主题 - 工具和来源 - 实体等);D. Wang 等[3]整合统计学特征、决策树和支持向量机算法以及已有知识来提取未知文本语义实体的新颖策略;再如 MnM[4]、OntOMat[5]以及 AKT(advanced knowledge technology)项目的 Melita[6],半监督类工具有 IBM 设计实现的 SemTag[7-8]、Armadillo[9]以及 Y. F. Guo[10]提出的最小监督学习方法用于医学文献的综述研究。这些方法都在各自的研究领域取得了一定的成效,其中 Y. F. Guo 提出的利用语篇修辞与词汇本身特征的最小监督学习方法对医学论文中的“信息结构”的识别,主要从研究背景、研究方法与研究结果的视角进行识别,只是“研究方法”的识别准确率仅仅达到 29%,召回率也只有 50%。

而苏牧、肖人彬等提出神经网络识别方法和宽度优先法可以将聚类后的各个语句进行知识形式转换,从而完成由自然语言问卷到面向对象知识体系的知识抽取过程[11];许勇、宋柔等提出一种基于隐马尔科夫模型的方法标注大百科全书,即利用知识点在条目文本中的转移规律以及知识点的词特征分布来判断每个句子的知识点类别[12]。另外,“基于本体标引文献的工具”(An Ontology Based Tool for Preparation of Articles)[13]项目组在 2007 年 -2009 年期间开展全文挖掘与标引工作中,抽象出“科技论文核心信息(core information scientific papers,CISP)”概念,该方法的实验结果相对较好,但是受到知识语料的限制。

3.2 基于规则模式匹配的方法

H. Hongbo[14]利用规则实现科技论文中所描述方法的抽取,准确率达到了 85%,但是未对其他类型指纹进行研究。C. D. Manning[15]利用信息抽取模式(information extraction patterns)实现技术方法以及分类主题短语的识别与抽取,准确率仅有 20%。D. Kiela[16]以及 Y. F. Guo[17-18]利用话题在科技论文中的属性规则实现研究方法的识别与抽取,属性包括位置、时态、动词、语法等。J. E. Kohler[19]使用基于指示词的规则实现期刊文献摘要的研究方法的识别,指示词如 method、analysis、algorithm、approach 以及 mode 等。刘一宁、郑彦宁等针对学术期刊设计了一种学术定义抽取系统,通过混合使用模式规则、语法规则和词频统计以达到定义抽取的目的[20]。丁君军、郑彦宁等对学术期刊中的属性描述进行了数量关系和情感信息的分析[21];以作战文书为代表的科技论文构造上,郭忠伟、周献中和黄志同等构造各类作战文书的 Schema 库,利用 Schema 上的修辞谓词抽取相应的知识,最终构造文书内容[22]。

3.3 基于网络协同编辑方法

SemLib EU project(2012)开发了 Pundit[23],满足用户在标注网页的同时构建结构化数据,支持群组用户分享标注和协同建立结构化知识,通过三元组存储和关系数据库实现语义标注对象的持久化存储。英国开放大学开发的 SWEET[24](semantic web services editing tool),提供了一个轻量级的 Web APIs 的语义标注 Web 应用,基于 JS 和 Ext GWT 实现,用户直接嵌入到 Web 浏览器中即可使用。

3.4 基于语法关系方法

S. Gupta[25]等提出使用句法依赖树实现科技论文中的使用到的技术知识点的标注与抽取。S. Bethard[26]利用语言学实现问答系统中的事件及其语义类型识别,取得了较高的准确率。ReVerb 语义标注系统[27]引入了语法和词汇限制,主要体现在动词表示的两元关系上,其效果比 TextRunner 和 WOE 等软件,无论在召回率还是准确率上都有显著提高。德国莱比锡大学 AKSW 研究组提出的 FOX[28](federated knowledge extraction framework)框架整合了关联数据云平台,利用 NLP 算法从自由文本中抽取 RDF 三元组,同时也整合了命名实体识别、关键字抽取以及语义关系抽取等工具。

综上所述,现有的技术方法对于特定研究环境中的语义知识识别具有一定成效,也为研究设计指纹识别提供了技术支撑,但总体来说仍具有较强的学科领域依赖性,对于无监督指导学习的环境适应性不够,在无领域知识组织体系(KOS)以及人工定义规则的前提下,无法开展科技论文的研究设计指纹识别的应用,而且定义规则对于专业人员要求较高,特别是“研究设计指纹”的识别。因此,本研究从科技论文本质出发,遵

循科技论文的写作指南规范、科技期刊 CHECKLIST 规程、科研实验的 CHECKLIST 规程、研究设计指纹的描述表达习惯等客观规律与现象，设计研究设计指纹识别模型算法。

4 研究设计指纹自动识别方法的设计与实现

结合科技论文写作指南以及组织结构的外部结构特征与内部内容特征因素，本研究提出“基于多规则模式混合机器学习方法”来识别科技论文的研究设计指纹，即在综合运用语义指示词规则、语义行为词规则、语义词序列对规则、篇章修辞规则、位置特征规则以及上下文指纹特征等多种规则模式混合基础上，首先基于已有本体库知识进行标注，之后使用机器学习方法来进一步学习与丰富规则模式知识库，尽可能更全面、更准确地识别与标引科技论文中的研究设计指纹，具体实施技术路线见图1，包括基于文献段落的研究设计线索自动发现、线索知识库的构建（线索规则库）、基于句子粒度的研究设计指纹自动识别与基于术语粒度的研究设计指纹自动识别，最终形成篇章的研究设计指纹知识库，实现科技论文的指纹识别。

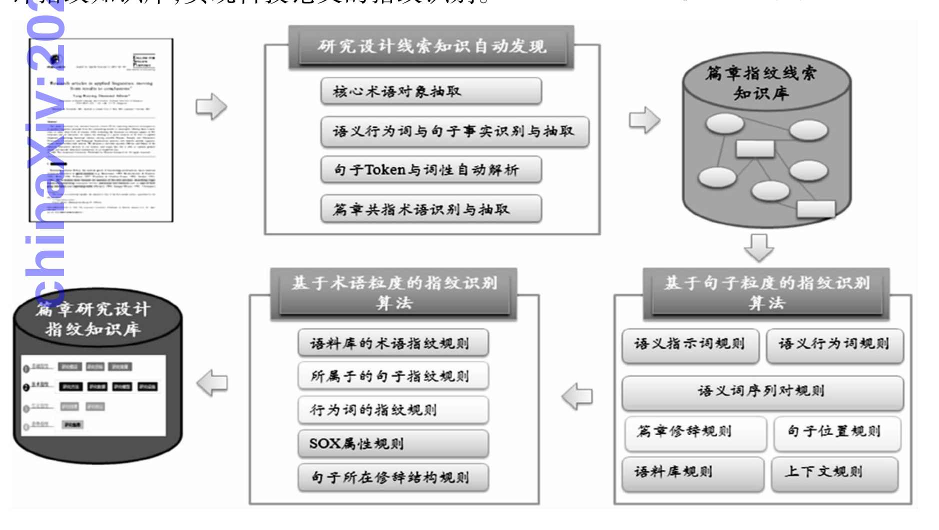


图1 研究设计指纹识别方法的实施技术路线

基于上述技术路线，算法设计共分6步，具体设计与实现描述如下：①基于文献段落的指纹线索发现与计算，包括分词、词性标注、命名实体识别、词根提取、共指词汇提取、核心词汇提取、核心语义行为提取以及句子事实提取；②删除噪音指纹对象数据（术语粒度的指纹对象）；③句子粒度的研究设计指纹特征识别；④术语粒度的研究设计指纹特征识别；⑤研究设计指纹识别结果的修正；⑥生产并创建研究设计指纹索引知识库。

4.1 基于文献段落的研究设计指纹线索自动发现方法

鉴于科技论文全文作为非结构化文本而较难识别指纹特征问题，笔者提出借助自然语言处理（NLP）的相关技术方法来实现科技论文全文的自动解析、知识重组与结构化表示，核心算法基于知识对象抽取和词特征抽取实现。

4.1.1 基于知识对象抽取的线索发现方法 知识抽取[29]是指从数字资源中识别、发现和提取出概念、类型、事实及其相关关系、约束规则，以及进行问题求解的步骤、规则的过程。本着这一指导思想，笔者采用 Stanford CoreNLP[30] 相关技术方法，实现从知识抽取过程中发现与抽取指纹特征线索，包括线索词和线索模式，例如词性标注、命名实体识别、分词解析器、语法分析、共指分析以及引导模式学习等功能方法，从术语抽取、语法分析以及事实抽取3个方面，设计从知识抽取过程中发现与抽取指纹特征线索的实现方法（见图2）。

（1）基于科技术语词的线索词发现。首先通过使用领域科技术语规范库，实现基于句子粒度的线索词抽取；其次，利用分词解析器实现基于句子粒度的自由术语的识别与抽取，同时结合识别出自由术语的词性规则，以连续词性的最大语义块原则，选择自由术语块作为备选线索词，最大范围地保障术语词的上下文背景；最后，利用术语相似度算法，即基于余弦算法[31]实现自由术语词与规范科技术语的相似度计算，以便进一步规范与标注自由术语词，其中在相似度的阈值上选择0.8，即 sim(x,y) >=0.8 时，计算公式如下公式（1）所示，其中 x 与 y 是表示两个术语词的向量，x = <x[1],…,x[m]>，y = <y[1],…,y[m]>。

$$sim(x,y)=\frac{\sum_{i=1}^{m}x[i]\cdot y[i]}{\|x\|\ \|y\|}=dot(x,y)=\sum_{i=1}^{m}x[i]\cdot y[i]$$

（2）基于句子语法的线索规则发现。利用自然语言处理技术，基于句子粒度进行 token 解析、词性标注

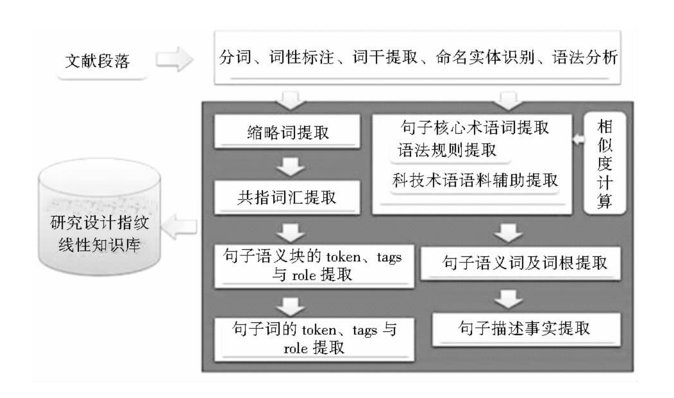


图 2 研究设计指纹识别线索发现与计算流程

以及最大语义块的提取，之后进行结构化存储。通过分析这些基于句子语法结构化的基础数据，形成系列的线索规则，来辅助识别指纹特征类型，例如，一个识别研究方法指纹特征类型的规则：(JJ | NN | NNS | NNP | NNPS) + (method | approach | measure | ···.)，根据该线索规则，基本可以判定语义块(JJ | NN | NNS | NNP | NNPS)为一个研究方法指纹。

(3)基于事实的线索模式发现。J. Bessin[32] 在基于大数据开展联邦业务分析中强调，事实抽取作为大数据分析的重要核心组织之一，其主要目标是确定重要描述的语义以及它们相互之间的关系。笔者将利用科技论文中"形成一个事实的行为动词"为研究切入点，发现与识别与该事实相关的主体、客体、动作、行为、状态、处所以及时间等事实属性，以发现研究设计指纹与事实行为的特征关系以及线索规则模式，即利用指纹规范语料库中的规范指纹特征词，与该事实相关的主体、客体、动作以及行为进行匹配，如果匹配成功，则该事实形成一个指纹线索规则模式，如果线索规则库中已存在，则不存储，否则直接存入到规则模式库中。例如句子：Web administrators, through the Robots Exclusion Protocol, use a special-format file called robots，通过事实抽取规则，则发现以下 3 个描述事实，即：

- Web administrators -》use -》special-format file
- Web administrators -》use -》Robots Exclusion Protocol
- ---》called -》robots

4.1.2 基于特征指示词的线索发现方法 在科技论文内容撰写过程中，研究设计指纹成果的描述具有一定规则的表述习惯以及依赖于上下文背景的普遍现象。基于这一规律现象，本研究提出基于特征指示词的线索发现方法，主要基于指示性名词、指示性行为词以及指示性共现词 3 种特征指示词来发现研究设计指纹线索。

(1)基于指示性名词的线索发现。通过一个名词术语即可标识一个知识对象的指纹特征类型，即定义为指示性名词线索，例如句子：The finite projective geometrymethod was first applied to determine the weight hierarchies，通过指示词 method 基本可以确定 The finite projective geometry 的指纹特征类型为"研究方法"。

(2)基于指示性行为词的线索发现。通过一个行为词即可标识一个知识对象的指纹特征类型，即定义为指示性行为词线索，例如句子：The finite projective geometry method was first applied to determine the weight hierarchies，通过指示词 applied 基本可以确定该句子的指纹特征类型为"研究方法"，而根据该句子的句法分析可知，applied 为被动语态，因此可以推断知识对象 The finite projective geometry 的指纹特征类型为"研究方法"。

(3)基于指示性共现词的线索发现。通过一个词对或多个单词或多个词组共现的现象即可标识一个知识对象的指纹特征类型，即定义为指示性共现词线索，例如句子：The finite projective geometry method was first applied to determine the weight hierarchies，通过指示词 method 和 applied 的共同出现在一个句子中，基本可以确定该句子的指纹特征类型为"研究方法"，因为在科技论文描述一个研究方法时，经常以"应用(apply)一个 XXX 方法(method)来解决 XXX 问题"。

4.2 基于句子粒度的研究设计指纹自动识别方法

4.2.1 算法设计与实现 句子粒度的研究设计指纹自动识别算法设计主要从语义指示词规则、语义行为词规则、语义共现词规则、句子所在段落位置、句子所属修辞类型、语料库规则及上下文指纹类型规则等多个规则模式来综合判断一个句子知识的最可能的研究设计指纹特征类型。其中，句子粒度的指纹特征向量(sentence fingerprinter space vector，简称 SSV)，主要以与句子知识单元相关的指纹特征类型相关因素为主要指标维度，包括句子的核心术语(coreterms)、位置等 10 个维度，即：SSV = (SentenseID, Text, CoreTerms, CorpusWords, CorpusWordsType, SectionType, Location, Action, ActionType, ActionTense)。

核心算法设计分为 2 个阶段：第 1 个阶段是基于语义指示特指词的构建算法，利用语法规则、定义规则以及 Be 动词规则来识别与标注指纹特征；第 2 个阶段是基于指示代词特征的构建算法，利用指示代词来识别与标注当前句子的指纹特征类型，同时建议了最邻

近上下文句子的指纹特征类型。

4.2.2 句子粒度研究设计指纹特征类型综合评判方法 投票方式是识别句子可能的指纹特征类型的主要方法，投票者代表一个类型的规则，每个规则都有投票权利，但是由于身份不同，所以权重不同，其权重按照研究设计的指纹识别模型权重值的分配规则执行（见表1）。如果每位投票者的得分是权重值*0（反对）或者权重值*1（赞同），那么一个句子所属一个指纹特征类型的最终得分为每位投票者得分总和，最终按照得分从高到低进行排序，最高者则被识别为最可能的句子指纹特征类型，投票得分的算法如下：

Sentence_FP_Score = 2 * IndicatingWordsValue + 1 * ActionWordsValue + 2 * Co-occurrenceValue + 0.5 * LocationValue + 0.5 * ORB-Value，其中 IndicatingWordsValue 表示句子中是否包括指示词的变量，如果包括 IndicatingWordsValue = 1，否则 IndicatingWordsValue = 0，其他变量的计算方法等同。

表1 指纹识别模型权重值分配层次

序号	分配层次名称	分值	描述
1	语义性特征权重值	2.0分	语义层面权重最大，直接从语义层面上标识知识单元的指纹特征类型
2	基准性特征权重值	1.0分	主要从语义行为词视角进行设置
3	强调性特征权重值	0.5分	如果满足某一条件，则指纹特征类型的强度就增加

4.3 基于术语粒度的研究设计指纹自动识别方法

4.3.1 算法设计与实现术语粒度的研究设计 指纹识别算法设计主要从语料库特征、所在句子指纹特征、所在句子行为词的指纹特征、SOX属性特征以及句子所在修辞结构特征5个方面来综合判定一个术语知识的最可能的研究设计指纹特征类型。其中，术语粒度的指纹特征向量（term fingerprinter space vector，简称TSV），主要以与术语知识单元相关的指纹特征类型相关因素为主要指标维度，包括术语是否是规范语料词（isCorpus）等9个维度，即：TSV =（Term，isCorpus，CorpusType，ParagraphType，Location，SentenceFP，Role，Action，ActionTense），其中 Term 为术语、isCorpus 为语料词标识、CorpusType 为语料词指纹类型、ParagraphType 为段落修辞类型、Location 为所在位置、SentenceFP 为所在句子指纹类型、Role 为术语角色、Action 为行为词、ActionTense 为行为词时态。

核心算法设计包括：第1步计算 CorpusWordsValue 得分，即使用语料库进行判断该术语词是否在语料库中，如果是，则该术语指纹类型+2分；第2步计算 SentenceFPValue 得分，即当前术语词识别判断的指纹特征类型是否与所属的句子指纹特征类型一致，如果一致，则该术语指纹类型+1分；第3步计算 ActionWordsValue 得分，即当前术语词所在句子的行为词，其具有的指纹特征类型是否与识别判断的指纹特征类型一致，如果一致，则该术语指纹类型+1分；第4步计算 SOXValue 得分，即当前术语词是否核心词汇，如果是，则该术语指纹类型+0.5分；第5步计算 ORB-Value 得分，即当前术语词识别判断的指纹特征类型是否与所在句子的修辞结构类型一致，如果是，则该术语指纹类型+0.5分。

4.3.2 术语指纹特征类型综合评判方法 术语粒度的指纹识别算法的设计与句子粒度的指纹识别算法设计基本类似，本质区别在于句子的指纹特征类型作为了术语粒度指纹识别的一个参数，即一个句子的指纹特征类型一定程度上影响着该句子中核心术语词的指纹特征类型的识别。所以仍然采取综合利用各个识别规则模式，使用投票的方式来识别术语知识可能的指纹特征类型，其中，投票者代表一个类型的规则，每一个都有投票的权利，但是由于其身份不同，所以投票决定的权重也不同。其中投票得分的算法如下（权重值参照表1）：

Term_FP_Score = 2 * CorpusWordsValue + 1 * SentenceFPValue + 1 * ActionWordsValue + 0.5 * SOXValue + 0.5 * ORB-Value，其中的变量解释参照4.3.1中的算法实现部分。

5 实验和结果

5.1 语料库与实验数据准备

5.1.1 语料库数据 材料构建语料库数据材料主要包括数据挖掘领域方向的专业术语、领域KOS、研究设计指纹特征指示词以及规则模式集，这些既是指纹特征类型识别的直接线索，也是机器学习的语料依据，用于指纹线索的发现与计算。具体创建过程为：①专业术语，该部分的语料主要使用"十二五"科技支撑计划项目成果，即科技知识组织体系（STKOS）[33]的工学人工智能方向的科技术语；②领域KOS，针对 Data Mining 研究方向，使用IEEE[34]叙词表[35]；③研究设计指纹特征指示词，采用 WordNet[36]（主要利用同义词以及词

典）、VerbNet[37]（主要利用行为词角色）、计算机科学研究论文的语料分析[38]以及期刊发表要求纲要、科技论文以及科技报告撰写的纲要等材料，构建指示词语料。本次实验创建研究设计指纹的线索指示性语料 235 个。

5.1.2　科技论文全文数据　材料的准备作为信息解决方案提供商，Elsevier 将科技论文的全文以“富媒体”HTML 格式进行结构化展示，有效支持了科研用户以“Play”模式来科学、合理地深度利用科技论文全文。处于上述良好的信息环境下，以主题词“Data Mining”为检索词，利用手工下载保存的模式，从 Elsevier 官方数据库下载 HTML 格式的科技论文全文文件，共计 100 篇（拟采用小样本数据，对本研究提出的研究方法体系的可行性进行验证分析，后续在该方法的普惠性方面将扩大数据量与实验领域进行对比验证），作为本研究研究方法可行性分析的原始数据。

5.2　实验结果评价分析

实现效果分别如图 5、图 6 与图 7 所示。同时，为了验证论文提出的研究设计指纹自动识别算法对 9 种研究设计指纹识的识别效果，采取领域专家对比分析法，邀请 10 位专家分别对 50 篇文献研究设计指纹识别结果进行判读，判读结果详见表 2，并通过专家认可度指数来标识各个区间的判读结果分布，专家认可度指数的计算公式为：认可度 = 票数 * 区间最低准确率/Sum（票数 * 区间最低准确率），其中 Sum（票数 * 区间最低准确率）为各个区间的最小值与该区间票数乘积的总和，专家认可度分析详见表 3。

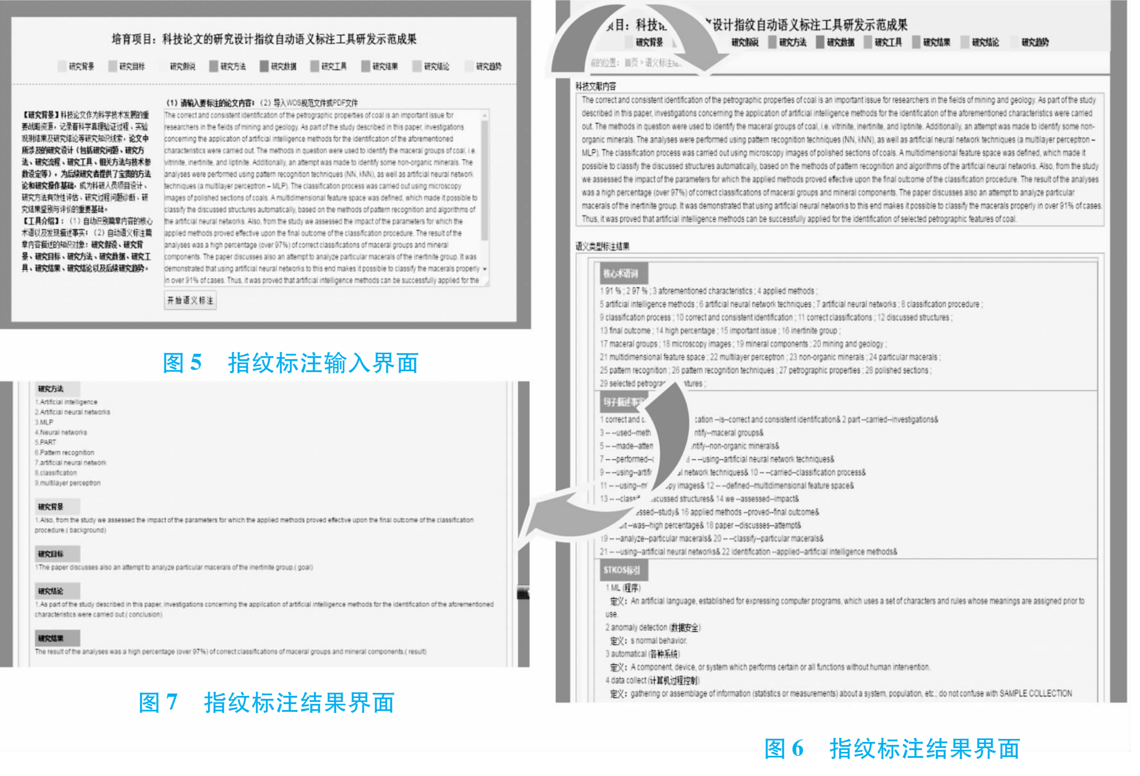


图 5　指纹标注输入界面

图 7　指纹标注结果界面

图 6　指纹标注结果界面

表 2　研究设计指纹识别结果专家判读结果

指纹类型	专家判读正确率的票数分布信息				专家判读覆盖率的票数分布信息			
	<60%	60% –70%	70% –80%	>80%	<60%	60% –70%	70% –80%	>80%
研究方法	–	–	5	45	–	–	–	50
研究工具	–	3	6	41	–	5	20	25
研究数据	5	10	15	20	10	10	25	15
研究假设	2	7	8	33	–	2	3	45
研究目标	–	–	7	43	–	–	2	48
研究背景	–	–	8	42	–	–	10	40
研究结果	–	2	9	39	–	–	3	47
研究结论	–	4	8	38	–	5	5	40
研究趋势	3	10	11	26	1	2	6	41

表 3　研究设计指纹识别结果的专家认可度分析

指纹类型	指纹识别结果准确率“认可度”分析				指纹识别结果覆盖率“认可度”分析			
	<60%	60%－70%	70%－80%	>80%	<60%	60%－70%	70%－80%	>80%
研究方法	－	－	9.0%	91.0%	－	－	－	100%
研究工具	－	4.7%	10.8%	84.5%	－	8.1%	37.8%	54.1%
研究数据	8.5%	16.9%	29.5%	45.1%	14.4%	14.4%	42.3%	28.9%
研究假设	3.3%	11.2%	15.0%	70.5%	－	3.1%	5.3%	91.6%
研究目标	－	－	12.5%	87.5%	－	－	3.6%	96.4%
研究背景	－	－	4.3%	85.7%	－	－	18.0%	82.0%
研究结果	－	3.1%	16.3%	80.6%	－	－	5.3%	94.7%
研究结论	－	6.3%	14.6%	79.1%	－	7.9%	9.0%	83%
研究趋势	6.7%	16.2%	20.8%	56.3%	1.7%	3.0%	10.8%	84.5%

表 3 的整体“认可度指数”结果显示，本研究提出的“基于科技论文的研究设计指纹识别方法”在准确率和覆盖率方面，基本达到了实验预期与目的，其中研究方法等 7 种特征指纹的识别准确率和覆盖率都达到了 80% 以上，而认可度指数相对较低的其他两种特征指纹的识别结果也主要分布在 70%－80% 的区间中。部分指纹类型的准确率或覆盖率相对较低，可能原因如下：①研究数据、研究假说与研究趋势指纹在科技论文中的描述特征不够明显，显性特征比起其他指纹特征相对较弱，例如：研究数据，提到 data 术语大多泛泛而指，不能确定具体的研究数据指纹；研究假说，科技论文中都会有提及，但是很多时候是通过笔者推理暗示来表达，一定程度上难从中识别；而研究方法、研究结论等指纹特征类型，表述的特征性相对较强，例如提出了 XXX 方法，最终得出了 YYY 结论等。②由于实验数据集相对较小，也影响了标引语料与特征规则模式的数量相对较少，一定程度上也影响了识别的准确率与覆盖率。

基于上述分析结论，对认可度相对较高的指纹类型，将基于目前指纹识别方法，进一步总结与发现相关规律、特征，提升其准确率与覆盖率；对认可度较低的指纹类型，将对其相关的计算指标与方法进一步调整，同时将利用深度学习，进一步更全面的、更细节性的挖掘与学习研究数据、研究假说和研究趋势 3 种指纹特征类型在科技论文全文中的描述特征。

6　结论与展望

本研究通过对科技论文的内容特征进行全面分析，提出了基于多规则模式混合机器学习的研究设计指纹自动识别算法，实证分析结果表明，该方法在大多数指纹特征类型的识别上效果显著，特别是研究方法、研究结论等指纹类型，有效地识别与抽取了隐含在科技论文全文中的重要指纹知识，但在某些指纹特征的识别上，比如研究数据，还需要进一步完善。

未来，笔者将进一步对影响研究设计指纹识别的相关因素进行全面分析，对评估研究设计指纹识别效果的有效性方法进一步改进，以“理工农医”不同的研究应用领域，开展更为广泛的应用示范，尽可能全面地发现指纹识别算法可能存在的问题，以提升与完善指纹识别算法模型的识别效果。

在此基础上，利用该指纹识别方法，对以期刊为核心的海量科技文献的元数据进行指纹识别，一方面构建学术界研究设计指纹之间丰富的关联关系，另一方面围绕专家、机构等科研实体，发现并构建出各自的研究设计指纹知识库，以增强学术知识计算的能力，提升基于大数据计算的知识发现的效果。

参考文献：

[1] 钱力，张晓林，王茜. 基于科技文献的研究设计指纹描述框架研究[J]. 大学图书馆学报,2015(1):14－20.

[2] GIRJU R, BEAMER B, ROZOVSKAYA A, et al. A knowledge-rich approach to identifying semantic relations between nominals [J]. Information processing & management an international journal, 2010, 46(5):589－610.

[3] WANG D, LIU X, LUO H, et al. A novel framework for semantic entity identification and relationship integration in large scale text data[J]. Future generation computer systems, 2016, 64(C):198－210.

[4] VARGASVERA M, MOTTA E, DOMINGUE J, et al. MnM: ontology driven semi-automatic and automatic support for semantic markup[C]// International conference on knowledge engineering and knowledge management. London: Springer-Verlag, 2002: 379－391.

[5] HANDSCHUH S, STAAB S, CIRAVEGNA F. S-cream——Semi-automatic CREAtion of metadata[C]//Knowledge engineering and knowledge management. Ontologies and the semantic Web. London: Springer-Verlag,2002:358－372.

[6] Advanced knowledge technologies[EB/OL]. [2017-09-26]. http://www.iam.ecs.soton.ac.uk/projects/akt/.
[7] CIRAVEGNA F, DINGLI A, PETRELLI D, et al. User-system cooperation in document annotation based on information extraction [C]// International conference on knowledge engineering and knowledge management. Ontologies and the semantic web. London: Springer-Verlag, 2002:122-137.
[8] DILL S, EIRON N, GIBSON D, et al. A case for automatedlarge scale semantic annotation. [EB/OL]. [2016-10-20]. http://www.websemanticsjournal.org/index.php/ps/article/viewFile/30/28.
[9] CIRAVEGNA F, CHAPMAN S, DINGLI A, et al. Learning to harvest information for the semantic web[C]//Proceedings of the 1st European semantic web symposium. Greece: Heraklion, 2004: 312-326.
[10] GUO Y F, SILINS I, STENIUS U, et al. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review[J]. Bioinformatics, 2013, 29(11):1440-1447.
[11] 苏牧,肖人彬. 基于语句聚类识别的知识动态提取方法研究[J]. 计算机学报,2001,24 (5) : 487-495.
[12] 许勇,宋柔. 基于 HMM 的百科辞典文本中句子的知识点分类[J]. 计算机工程与应用,2005:41(4):35-38.
[13] SOLDATOVA L N, LIAKATA M. An ontology methodology and CISP-the proposed core information about scientific papers[EB/OL]. [2016-09-24]. https://www.aber.ac.uk/en/media/departmental/impacs/computerscience/pdfs/ReportCISPshort.pdf.
[14] HOUNGBO, HOSPICE, MERCER R E. Method mention extraction from scientific research papers[C]//24th International conference on computational linguistics - proceedings of COLING 2012. New York: Curran associates, 2012.
[15] GUPTA S, MANNING C D. Analyzing the dynamics of research by extracting key aspects of scientific papers[C]//Proceedings of 5th international joint conference on natural language processing. New York: Curran associates, 2011: 1-9.
[16] KIELA D, GUO Y, STENIUS U, et al. Unsupervised discovery of information structure in biomedical documents[J]. Bioinformatics, 2015, 31(7): 1084-1092.
[17] GUO Y F, SILINS I, STENIUS U, et al. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review[J]. Bioinformatics, 2013, 29(11): 1440-1447.
[18] GUO Y F, REICHART R, KORHONEN A. Improved information structure analysis of scientific documents through discourse and lexical constraints[C]// Proceedings of NAACL-HLT, Association for Computational Linguistics. New York: Curran associates, 2013: 928-937.
[19] ECKLE-KOHLER J, NGHIEM TD, GUREVYCH I. Automatically assigning research methods to journal articles in the domain of social sciences[J]. Proceedings of the American Society for Information Science and Technology, 2013, 50(1): 1-8.
[20] 刘一宁,郑彦宁,化柏林. 学术定义抽取系统实现及实验分析[J]. 情报理论与实践,2011,34 (12): 15-19.
[21] 丁君军,郑彦宁,化柏林. 基于规则的学术概念属性抽取[J]. 情报理论与实践,2011, 34 (12): 10-14,33.
[22] 郭忠伟,周献中,黄志同. 作战文书自动生成系统中内容规划的设计[J]. 火力与指挥控制,2002,27 (4) : 51-54.
[23] Pundit - Semantic annotation tool[EB/OL]. [2017-03-20]. http://thepund.it/.
[24] SWEETEB[EB/OL]. [2017-03-20]. http://sweet.kmi.open.ac.uk/.
[25] GUPTA S, MANNING C D. Identifying focus, techniques and domain of scientific papers[EB/OL]. [2017-03-20]. https://www.researchgate.net/publication/267232558_Identifying_Focus_Techniques_and_Domain_of_Scientific_Papers.
[26] BETHARD S, MARTIN J H. Identification of event mentions and their semantic class[C]//Proceedings of the 2006 conference on empirical methods in natural language processing. Sydney: Emnlp, 2006: 146-154.
[27] FADER A, SODERLAND S, ETZIONI O. Identifying relations for open information extraction[C]//Conference on empirical methods in natural language processing. Edinburgh: Association for computational linguistics, 2011:1535-1545.
[28] FOX - Agile knowledge engineering and semantic web (AKSW) [EB/OL]. [2017-03-20]. http://aksw.org/Projects/FOX.html.
[29] 张智雄,吴振新,刘建华,等. 当前知识抽取的主要技术方法解析[J]. 现代图书情报技术, 2008, 24(8): 2-11.
[30] MANNING C D, SURDEANU M, BAUER J, et al. The stanford-corenlp natural language processing toolkit[C]// Proceedings of 52nd annual meeting of the Association for Computational Linguistics: system demonstrations. Maryland: Curran associates, 2014: 55-60.
[31] LEE D, PARK J, SHIM J, et al. An efficient similarity join algorithm with cosine similarity predicate[C]//International conference on database and expert systems applications. Heidelberg: Springer, 2010: 422-436.
[32] BESSIN J, DAS A. Big data analytics federal business analytics. [EB/OL]. [2017-03-20]. https://www.xerox.com/downloads/services/white-paper/big-data-analytics.pdf.
[33] 孙坦, 刘峥. 面向外文科技论文信息的知识组织体系建设思路[J]. 图书与情报,2013(1):2-7.
[34] IEEE 互动百科[EB/OL]. [2017-04-10]. http://www.baike.com/wiki/IEEE.
[35] IEEE_thesaurus_2013. [EB/OL]. [2017-04-10]. https://www.ieee.org/documents/ieee_thesaurus_2013.pdf.
[36] About WordNet[EB/OL]. [2017-03-20]. http://wordnet.princeton.edu/.

[37] Martha Palmer[EB/OL]. [2017－03－20]. http://verbs. colorado. edu/~mpalmer/projects. html.

[38] POSTEGUILLO S. The schematic structure of computer science research articles[J]. English for specific purposes, 1999, 18(2): 139－160.

作者贡献说明：

钱力：负责论文内容的撰写和论文修改；

张晓林：负责论文内容设计与审核；

王茜：负责论文修改。

Building and Implement on Automatic Identification Method of Research Design Fingerprint of Scientific Papers

Qian Li[1] Zhang Xiaolin[1] Wang Qian[2]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190

[2] Institute of Medical Information / Medical Library, CAMS&PUMC, Beijing 100005

Abstract: [**Purpose/significance**] Automatic identification and extraction of research design fingerprint from scientific papers is able to provide researchers with significant methodology and research support for project design, validity evaluation of research methods, problem diagnosis of research process and identification and evaluation of research results. [**Method/process**] The paper, based on the concept model of research design fingerprint in scientific papers, proposes amulti-rule hybrid machine learning methods to design and implement the fingerprint identification algorithm model and analyze and verify the feasibility and validity of the methodby sample data in the field of datamining. [**Result/conclusion**] The results show that in addition to the research data and research trends, the recognition accuracy of other research design fingerprint is almost 80%. And the acceptance of coverage, in addition to research tools and research data, is almost 80%.

Keywords: research design fingerprint semantic annotation knowledge extraction machine learning

